Clustering is a well-known task that has been used for decades. The idea is to take a set of items and group them into a number of clusters based on a similarity measure. K-means proposed in 1957 by Stuart Lloyd is one of the most widely used clustering algorithm and is still used today for its simplicity and efficiency. K-means has two main parts to clustering, the initial seeding process and the iteration process. The seeding process picks k initial seeds as cluster centres, and highly affects the accuracy of the final result in the algorithm. The iteration process dominates running time to move the centres around until it converges to an optimum. In this paper, we discuss a new method of the seeding process that gives us more accurate seeds to start the algorithm. We also discuss a novel approach to find an approximation of the correct number of clusters for a given dataset.

Notes:

Clustering is not a learning task

Add definition using objective function. Kmeans is a popular approach because of a reasonably fast heuristic to find the clusters based on the lyod algorithm and more recent achievments.

aspell -c -t

1. What percentage of clusters were classified successfully?
2. How many did I incorrectly classify against the ground truth.
3. Accuracy of kmeanspp and ask if there are statiscally significant differences between them?
4. If the numbers are very differnet, then we should run t-test?
5. Compare the objective function. Report the means and std deviation. Only compare means, then only compare standard deviation. Compare minimum as well as mean.
6. One sided standard deviation from the minimum (Ask Sarhan for this)
7. Look at 90% of the values from the minimum. Look at other percentages as well.
8. Are we buying robustness at the cost of the true optimum, the best result?